High-dimensional forgeries for fun and profit

A study of "A knockoff filter for high-dimensional selective inference"
Statistics 588 - Data Mining

Don Walpola

In many contemporary statistical applications, variable selection is a critical component of the analysis. Variable selection is concerned with determining the optimal, in some sense, subset of available features with which to model the desired target variable. For instance, in standard linear regression, we are concerned with models of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and typically $\boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. The columns of the design matrix $\mathbf{X}$ represent the p variables under consideration to model the response $\mathbf{y}$. But rather than use all p features of $\mathbf{X}$, we may believe that only some smaller set of $p_0 < p$ variables are truly associated with the response $\mathbf{y}$; or in the case that $p > n$, we may seek to find the set of $p_0 < n$ variables most strongly associated with $\mathbf{y}$ so that we can reduce our model to the identifiable case and perform regression. This is in contrast to classical statistical procedures, where inference (in the form of hypothesis testing or confidence estimates) on a pre-specified, fixed model was paramount. Issues similar to those of the classical setting must still be dealt with, however: is there a way to quantify how many of the selected variables are truly relevant features? Can we find any guarantees that the variables we select are not merely exhibiting spurious associations with the response?

1 Problem Description

One way to measure the effectiveness of a variable selection procedure is by the *False Discovery Rate*. Let $S = \{1, 2, \ldots, p\}$ be the index set of variables under consideration, $S_0 \subset S$ the index set of variables actually associated with the response $\mathbf{y}$, and $\widehat{S}_0 \subset S$ the estimated index set of relevant variables produced by our variable selection procedure. The false discovery rate is defined as the expected ratio of the number of irrelevant variables we select to the overall number of variables in our estimated set:

$$\text{FDR} = \mathbb{E}\left[\frac{\left|\{j \in S : j \in \widehat{S}_0 \text{ and } j \notin S_0\}\right|}{\max(|\widehat{S}_0|, 1)}\right] \tag{2}$$

The quantity inside the expectation is called the *false discovery proportion*, and the max function in the denominator guarantees that the ratio is defined even in the case where no variables are selected and $|\widehat{S}_0| = 0$. For the particular case of a linear regression model given by equation (1), the false discovery rate can be rewritten as

$$\text{FDR} = \mathbb{E}\left[\frac{\left|\{j \in S : \widehat{\beta}_j \neq 0 \text{ and } \beta_j = 0\}\right|}{\max(|\widehat{S}_0|, 1)}\right] \tag{3}$$

Note that $\widehat{S}_0 = \{j \in S : \widehat{\beta}_j \neq 0\}$ for the linear regression model.
In the original formulation of the knockoff filter [1], the false discovery rate is controlled by constructing a set of p false, or 'knockoff', variables $\widetilde{\mathbf{X}}$ that replicate the covariance structure of the original variables $\mathbf{X}$ both among themselves and between each knockoff $\widetilde{\mathbf{x}}_j$ and distinct original variable $\mathbf{x}_k$ for $k \neq j$, while each knockoff variable $\widetilde{\mathbf{x}}_j$is constructed to have correlation with its original version $\mathbf{x}_j$ reduced by a quantity s_j. Then a set of p statistics that satisfy two properties, which will be described below, is calculated for each pair of original and knockoff variables. Finally, based on the desired false discovery rate, a cutoff threshold is computed from the data in order to select a set of variables guaranteed to not exceed the desired FDR. We now give a more detailed description following the notation used in [1, 2].

Knockoff Procedure

1. **Construct 'Knockoffs'**: Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the data matrix, with the columns standardized to have unit Euclidean norm: $||\mathbf{x}_j||_2^2 = 1$ for all columns $1 \leq j \leq p$.

   The knockoff variables will be denoted $\widetilde{\mathbf{X}} \in \mathbb{R}^{n \times p}$, and must satisfy the following properties:

$$\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \mathbf{\Sigma} = \mathbf{X}^T\mathbf{X}$$
$$\mathbf{X}^T\widetilde{\mathbf{X}} = \mathbf{\Sigma} - \text{diag}(\mathbf{s})$$

   where $\text{diag}(\mathbf{s})$ is a diagonal matrix with the entries of the vector $\mathbf{s}$ along its diagonal. The vector $\mathbf{s}$ has components s_j that quantify the reduction in correlation between an original variable and its knockoff. As described by Barber and Candès [1], one method for constructing such a knockoff matrix $\widetilde{\mathbf{X}}$ is to select a vector $\mathbf{s} \in \mathbb{R}^p_{>0}$ with all positive components such that $\text{diag}(\mathbf{s}) \preceq 2\mathbf{\Sigma}$. Note that the partial order $\preceq$ is the *Loewner order* on the convex cone of real, symmetric positive semi-definite matrices S_p^+:

$$\text{For } \mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}, \quad \mathbf{A}^T = \mathbf{A} \text{ and } \mathbf{B}^T = \mathbf{B}$$
$$\mathbf{A} \preceq \mathbf{B} \iff \mathbf{B} - \mathbf{A} \text{ is positive semi-definite}$$

The knockoffs are then constructed as

$$\begin{aligned}
&\widetilde{\mathbf{X}} = \mathbf{X}(I - \Sigma^{-1}\text{diag}(\mathbf{s})) + \widetilde{\mathbf{U}}\mathbf{C} \\
&\text{where} \quad \widetilde{\mathbf{U}} \in \mathbb{R}^{n \times p} \\
&\text{such that} \;\; \text{span}(\widetilde{\mathbf{U}}) \perp \text{span}(\mathbf{X}) \text{ and } \;\; \widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} = \mathbf{I} \\
&\text{and} \quad \mathbf{C} \in \mathbb{R}^{p \times p} \\
&\text{such that} \;\; \mathbf{C}^T\mathbf{C} = 2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})
\end{aligned}$$

The product $\mathbf{C}^T\mathbf{C}$ is a Cholesky decomposition, which is guaranteed to exist by the condition $\mathbf{diag}(\mathbf{s}) \preceq 2\boldsymbol{\Sigma}$.

2. **Calculate statistics W_j for each pair of original and knockoff features $\mathbf{X}_j$ and $\widetilde{\mathbf{X}}_j$**: The statistics W_j must satisfy the two properties of *sufficiency* and *anti-symmetry*.

$$\begin{aligned}
&\textit{Sufficiency}: \quad \mathbf{W} = f([\mathbf{X}, \widetilde{\mathbf{X}}]^T[\mathbf{X}, \widetilde{\mathbf{X}}], [\mathbf{X}, \widetilde{\mathbf{X}}]^T\mathbf{y}) \\
&f : S^+_{2p} \times \mathbb{R}^{2p} \to \mathbb{R}^p \\
&\textit{Antisymmetry}: \quad \forall S' \subset S, \\
&W_j([\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(S')}, \mathbf{y}) = W_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}) \cdot (-1)^{\mathbb{1}\{j \in S'\}} \\
&\text{where } (V_{\text{swap}(S')})_j = \begin{cases} V_j & j \notin S' \\ V_{j+p} & j \in S' \end{cases} \quad (V_{\text{swap}(S')})_{j+p} = \begin{cases} V_{j+p} & j \notin S' \\ V_j & j \in S' \end{cases}
\end{aligned}$$

Verbally, sufficiency means that the vector of statistics $\mathbf{W} \in \mathbb{R}^p$ depends on the concatenated data and knockoff matrix $[\mathbf{X}, \widetilde{\mathbf{X}}]$ and response vector $\mathbf{y}$ only through the covariance $[\mathbf{X}, \widetilde{\mathbf{X}}]^T[\mathbf{X}, \widetilde{\mathbf{X}}]$ and the marginal covariance $[\mathbf{X}, \widetilde{\mathbf{X}}]^T\mathbf{Y}$. Antisymmetry means that swapping the columns of a variable with its knockoff in $[\mathbf{X}, \widetilde{\mathbf{X}}]$ changes the sign of the associated statistic W_j for that pair.

3. **Compute a data dependent threshold for the statistics**: For q the desired FDR, $t > 0$, and $\mathcal{W} = \{|W_j| > 0 : j \in S\}$ the set of non-zero statistics W_j, compute the threshold value:

$$T = \min\Big\{t \in \mathcal{W} : \frac{1 + |j : W_j \leq -t|}{\max(|j : W_j \geq t|, 1)} \leq q\Big\}$$

That is, T is the smallest nonzero W_j such that the ratio of 1 plus the number of statistics below $-T$ to the number of statistics above T does not exceed the target FDR. The max function in the denominator again ensures this quantity is always defined. The other technicality of the extra 1 added to the numerator allows some minor algebraic manipulations to make the threshold T directly control the FDR - without it, a modified version of the FDR with an extra term q^{-1} in the denominator is what ends up being controlled. The cost for this direct FDR control is a slightly more conservative threshold value. T is then used to select a set of

variables, with a guarantee that the FDR does not exceed q. Note that Barber and Candès distinguish these two threshold values; the T computed from the numerator inflated by 1 is referred to as *knockoff+*, while the T derived without augmenting the numerator and which only controls the modified FDR is named the *knockoff* procedure. We ignore this minor complication, and any following reference to the knockoff procedure here will be in reference to what is called *knockoff+* in [1].

There are several limitations to this procedure, some of which are addressed in [2]. One limitation is that the geometrical construction of the knockoffs requires that $n \geq 2p$ (or with some augmentations, $n \geq p$). This prevents the knockoff filter from being applied to 'high-dimensional' problems where $p > n$. Of course, this high-dimensional setting is arguably where variable selection is most often needed and applied.
A second limitation is that controlling the FDR is only practically useful in the sparse setting, where $|S_0| \ll |S|$. In the non-sparse setting, a more informative quantity to control would be the number of directional errors - that is, predicting $\widehat{\beta}_j > 0$ when in fact $\beta_j \leq 0$, or $\widehat{\beta}_j < 0$ when in fact $\beta_j \geq 0$. This leads to a definition of *directional false discovery rate*:

$$\mathrm{FDR}_{\mathrm{dir}} = \mathbb{E}\left[\frac{\left|\{j \in \widehat{S}_0 : \mathrm{sign}(\widehat{\beta}_j) \neq \mathrm{sign}(\beta_j)\}\right|}{\max(|\widehat{S}_0|, 1)}\right] \tag{4}$$

Recall that the sign function takes values in the set $\{-1, 0, 1\}$, with $\mathrm{sign}(0) = 0$. Thus, as defined, controlling the directional false discovery rate also controls the FDR because of the weak inequalities on the true parameter values β_j, and so $\mathrm{FDR}_{\mathrm{dir}}$ is also a sensible measure in sparse problems. A directional error is made when either an irrelevant variable is selected, or a relevant variable is selected but with the incorrect direction of effect. Note, however, that no distinction is made between these two types of errors by the $\mathrm{FDR}_{\mathrm{dir}}$.

2 Models, Methods, Algorithms

Barber and Candès demonstrate in [2] that, under the assumption of normally distributed observations, the knockoff filter can be used to control not just the FDR, but the $\mathrm{FDR}_{\mathrm{dir}}$. The subtleties and consequences of this assumption will be discussed in section 4. To explicitly address the case that $p > n$, Barber and Candès [2] propose a two stage procedure:

1. **Data splitting and initial screening:** From the original, full data set $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$, select a subsample $(\mathbf{X}^{(0)}, \mathbf{y}^{(0)}) \in \mathbb{R}^{n_0 \times p} \times \mathbb{R}^{n_0}$ with $n_0 < n$. The remaining data contain $n_1 = n - n_0$ samples. Using some variable selection procedure, select a subset $\widehat{S}_0 \subset S$ of features such that $|\widehat{S}_0| < n_1$.

2. **Apply Knockoff Filter on reduced model:** Using the remaining $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}) \in \mathbb{R}^{n_1 \times p} \times \mathbb{R}^{n_1}$ data points, reduce the model to contain only the features selected in

the previous step: $(\mathbf{X}^{(1)}_{\widehat{S}_0}, \mathbf{y}^{(1)}) \in \mathbb{R}^{n_1 \times |\widehat{S}_0|} \times \mathbb{R}^{n_1}$, and run the knockoff procedure as previously described, but with statistics W_j that take into account information about the signs of $\widehat{\beta}_j$ and $\widetilde{\beta}_j$.

One thing to note is that this two stage procedure for high-dimensions must now deal with issues of conditional inference in addition to the simultaneous inference concerns of the original knockoff filter in low-dimensions. So more, and different types, of the potential pitfalls of selective inference must be dealt with [5].
In addition to issues from selective inference, the loss in statistical power associated with all data-splitting procedures is of concern. The smaller size of the subsample $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ typically means a lowered ability to detect variables in S_0 compared to using the full dataset $(\mathbf{X}, \mathbf{y})$. The naive attempt to use $(\mathbf{X}_{\widehat{S}_0}, \mathbf{y})$ would destroy the ability of the knockoff filter to control the FDR [2]. Barber and Candès propose two surprising manipulations in order to regain some of the statistical power lost from data splitting:

1. *Data Recycling*: Reuse $\mathbf{X}^{(0)}_{\widehat{S}_0}$ when constructing the knockoff variables $\widetilde{\mathbf{X}}_{\widehat{S}_0}$. This differs from the naive approach in that we do not construct knockoffs for $\mathbf{X}^{(0)}_{\widehat{S}_0}$. The knockoffs $\widetilde{\mathbf{X}}^{(1)}_{\widehat{S}_0}$ are constructed as usual, but we set $\widetilde{\mathbf{X}}^{(0)}_{\widehat{S}_0} = \mathbf{X}^{(0)}_{\widehat{S}_0}$, so that we have $\widetilde{\mathbf{X}}_{\widehat{S}_0} = \begin{bmatrix} \mathbf{X}^{(0)}_{\widehat{S}_0} \\ \widetilde{\mathbf{X}}^{(1)}_{\widehat{S}_0} \end{bmatrix}$.
2. *Signed Statistics from Screening*: While performing the initial variable selection procedure on the data split $(\mathbf{X}^{(0)}, \mathbf{y}^{(0)})$, record not just the magnitude, but also the signs of the estimated effect, $\widehat{\text{sign}}^{(0)}_j$, of each feature $\mathbf{X}^{(0)}_j$ on $\mathbf{y}^{(0)}$. These initial estimates will be used to restrict the W_j used in the knockoff filter in order to achieve the desired directional control on the reduced model.

It may seem somewhat surprising that either of these manipulations should lead to any significant increase in statistical power, but Barber and Candès provide both empirical and heuristic justifications for their use. In brief, the intuition behind data recycling is the following: since the formulation of the knockoff estimate of the FDR relies on relevant features more likely to produce large $|W_j|$ than irrelevant features. Recycling the first n_0 data points shifts W_j to the right, as will be described below. If the relevant features are shifted to a greater degree than null features, then we should see an increase in power - and this differential shifting is justified by distributional arguments. The justification for using the estimated signs is even more hand-wavy - as long as we believe the estimate $\widehat{\text{sign}}^{(0)}_j$ does in fact indicate the more likely direction of the effect of $\mathbf{X}_j$, then the use of $\widehat{\text{sign}}^{(0)}_j$ for directional control has merit. Of course, this depends on the variable selection procedure used in initial screening to correctly include relevant variables and assess the direction of their effects, but typically used procedures such as the LASSO are usually believed able to do so. Care must be taken, however, as using the LASSO for initial variable screening is not guaranteed to achieve consistent recovery of the true

set of relevant variables S_0 unless certain conditions restricting the covariance between relevant and irrelevant features are satisfied [6, 7]. Even modifications such as the elastic net may incorrectly estimate the signs of selected features [7]. Such concerns are further exacerbated by the possibility that relevant features were not even in the initially considered set of p features in $\mathbf{X}$, as may occur for example when causal mutations have not been typed in genome data, but mutation sites located nearby on the chromosome to, and thus correlated with, the causal mutations *are* included in the observed data [2]. Failure to recover the true support of $\boldsymbol{\beta}$ will lead to bias in the estimates $\widehat{\boldsymbol{\beta}}_{\widehat{S}_0}$, which may possibly induce directional errors. Thus the performance of the selection procedure used in the screening step is of significant concern, and will be elaborated upon in section 4.

3 Empirical Evaluations

The ability of the knockoff filter to control the FDR_{dir} is tested on randomly generated simulated data. Barber and Candès demonstrate the knockoff filter on a simulated autoregressive model and a simulated response model derived from actual genome data [2]. In order to test the assertion (discussed in section 4 of [2]) that FDR_{dir} is controlled by the knockoff filter for normally distributed $\mathbf{X}$ with arbitrary covariance matrix, we generate a random covariance matrix for the multivariate normally distributed $\mathbf{X}$. In order to generate an arbitrary covariance matrix that is not dominated by its diagonal entries, we use the following procedure that will generate a symmetric, positive definite matrix:

- First set the number of desired variables, p, and the number of 'factors' $1 \leq k \leq p$ to generate the covariance matrix. The closer k is to p, the closer the resulting covariance matrix will be to a diagonal matrix. Conversely, the closer k is to 1, the less like a diagonal matrix the resulting covariance matrix will be.
- Generate a $p \times k$ matrix $\mathbf{W}$ with entries drawn from a uniform distribution. We use Uniform$[-10, 10]$ here. Also generate a $p \times p$ diagonal matrix $\mathbf{D}$ with positive diagonal entries drawn from a random distribution. We use Uniform$[1, 10]$ here.
- Compute the covariance matrix according to the formula

$$\boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}' + \mathbf{D}$$

Clearly $\boldsymbol{\Sigma}$ is symmetric, as $\mathbf{W}\mathbf{W}'$ and $\mathbf{D}$ are symmetric. The matrix $\mathbf{W}$ can be thought of as containing k 'basis' vectors, as the randomly generated entries will almost certainly produce linearly independent vectors. Now this product will have at most k non-zero eigenvalues, but these eigenvalues will be non-negative since standard linear algebra results shows that $\mathbf{W}'\mathbf{W}$ is positive semi-definite, and $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}\mathbf{W}'$ have the same non-zero eigenvalues. Although this basis is not orthogonal like an eigenbasis, it is diagonalizable and adding the diagonal matrix will make sure that all the eigenvalues of $\boldsymbol{\Sigma}$ will be positive. Thus we have produced a symmetric positive-definite matrix, which we may now use as a covariance matrix. The following heat map indicates the relative

magnitudes of a 500×500 matrix generated in this manner with $k = 20$ and obviously $p = 500$.

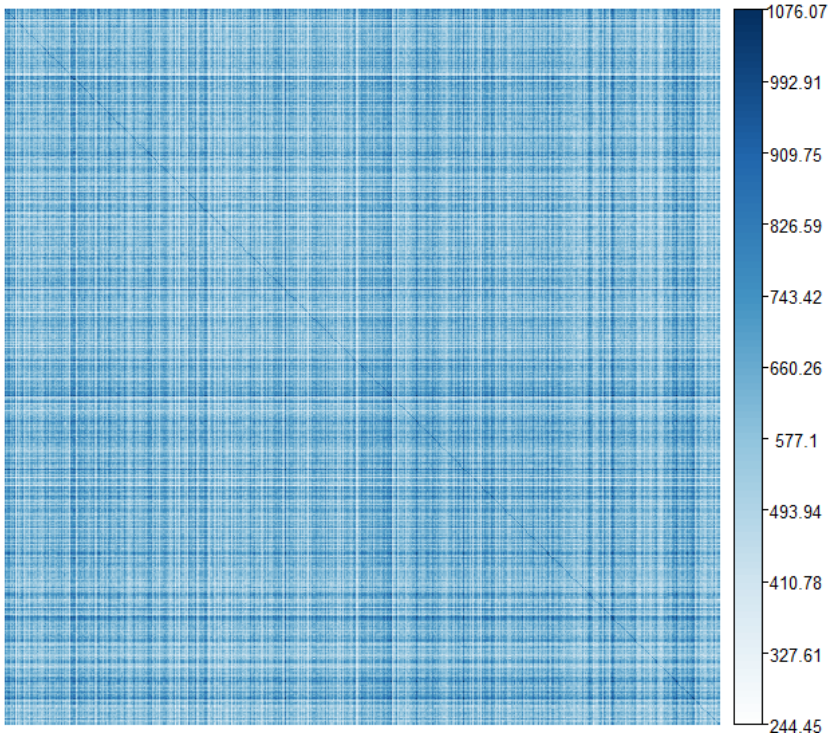

Figure 1: Heatmap of randomly generated 500×500 covariance matrix using $k = 20$ random 'factors' with entries drawn uniformly from $[1, 10]$.

The image was generated using the `corrplot` function in the `clusterGeneration` package in `R`. We can also gain a sense of the distribution of magnitudes by examining a histogram of the values of the matrix, displayed in Figure 2. The maximum entry in the matrix is 1164.36 and the minimum is 244.42.

A consequence of this construction is that there tends to be one principal component that accounts for the vast majority of the variance; the next 18 to 20 are substantially smaller, and the remaining eigenvalues of $\mathbf{\Sigma}$ are negligible in magnitude compared to the first k. The scree plot in figure 3 shows the variances associated with the first 50 principal components, demonstrating this behavior. While this would be considered good behavior to have for certain forms of dimension reduction, this may lead to a high level of covariance between important and irrelevant variables.

Using this covariance matrix $\mathbf{\Sigma}$, we now generate the data matrix $\mathbf{X}$ using a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. As we are simulating a high-dimensional data set with fewer samples n than variables p, we generate $n = 400$ samples and the matrix $\mathbf{X}$ is then 400×500 with each row an independent draw from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Once this is done, we normalize the data matrix so that each column has unit L_2 norm. We then randomly assign $p_0 = 50$ of the 500 variables to be relevant for our model, by generat-

]

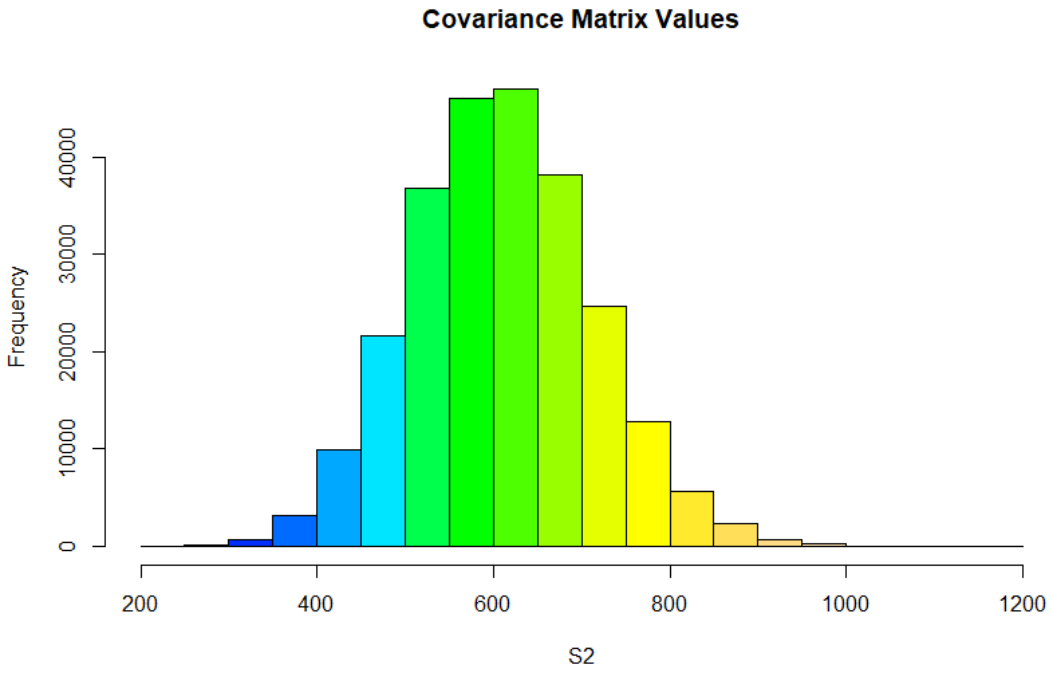

Figure 2: Histogram of randomly generated covariance matrix entries

ing a parameter vector $\boldsymbol{\beta}$ of length 500 and randomly selecting 50 of its indices to be non-zero. The magnitude of the non-zero entries β_{p_0} are all equal, but the signs are randomly distributed using a symmetric Bernoulli distribution. Also called a Rademacher distribution, it takes the values -1 and 1 with equal probability.

Using the generated design matrix, we now simulate a response according to the equation $y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. In order to assess the average performance, we apply the two-stage knockoff procedure for 100 trials using the current magnitudes for the non-zero β_{p_0}. At each trial the noise vector $\boldsymbol{\epsilon}$ is newly generated and the subsample n_0 used for initial screening is randomly drawn. The initial variable screening is performed using LASSO regression as implemented by the `glmnet` function in `R`. The knockoff statistic used during the second stage is the difference in magnitudes of LASSO regression coefficients in the reduced, screened model, $|\hat{beta}_j| - |\tilde{\beta}_j|$. Several functions in the `knockoff` package were used to perform the filtering at the second stage. To assess the performance for various coefficient magnitudes, we repeat this whole procedure for a range of signal magnitudes $|\beta_{p_0}|$. The results are plotted against the parameter values in figure 4. We see that the directional FDR is fairly consistent, and below the set level of 0.1 used for the simulations. Directional and standard measures of power are the same, supporting the claim that the knockoff filter controls directional error rates without modification. Power was measured as the fraction of true signals finally selected

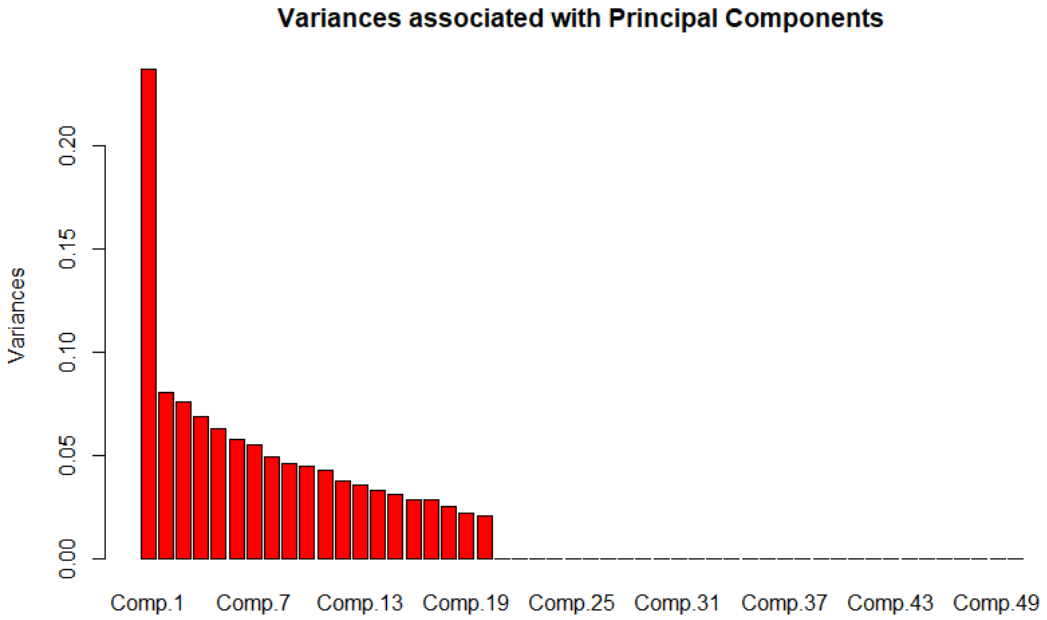

Figure 3: Scree plot of variances associated with principal components of random covariance matrix generated using $p = 500$ and $k = 20$.

after the second stage and application of the knockoff filter. Directional power further imposes that the signs of the estimate and true parameter must agree. The restricted versions considered only the true signals that made it past the initial screening stage, before the knockoff filter was applied in the second stage. The smaller denominator obviously leads to a larger magnitude for the restricted power. Weaker signal strengths are usually associated with a lower power, though it does not appear to be monotonic as a function of increasing parameter magnitude in the simulations performed. See the attached code in the file `Project Code.R` for details of the implementation.

As the average correlations between covariates are very high using the method above to generate the covariance matrix, we repeat our assessment using a slightly different covariance matrix. The only difference here is that we now take the weighted sum of two positive-semidefinite matrices as our covariance. The first matrix is generated exactly as above, while the uniform distribution used to generate the 'factor' matrix $\mathbf{W}$ for the second matrix now takes values over the symmetric range $[-10, 10]$. The covariance matrix we use is then $\mathbf{\Sigma} + 2\mathbf{\Sigma}_{\text{sym}}$, with twice the weight on the matrix generated using the symmetric uniform. The resulting matrix has a lower average value for the off-diagonal terms. Unlike the earlier covariance matrix, however, some of the entries here are negative, producing a qualitatively different covariance structure.

The results show that the dirFDR control is maintained, while all the average power

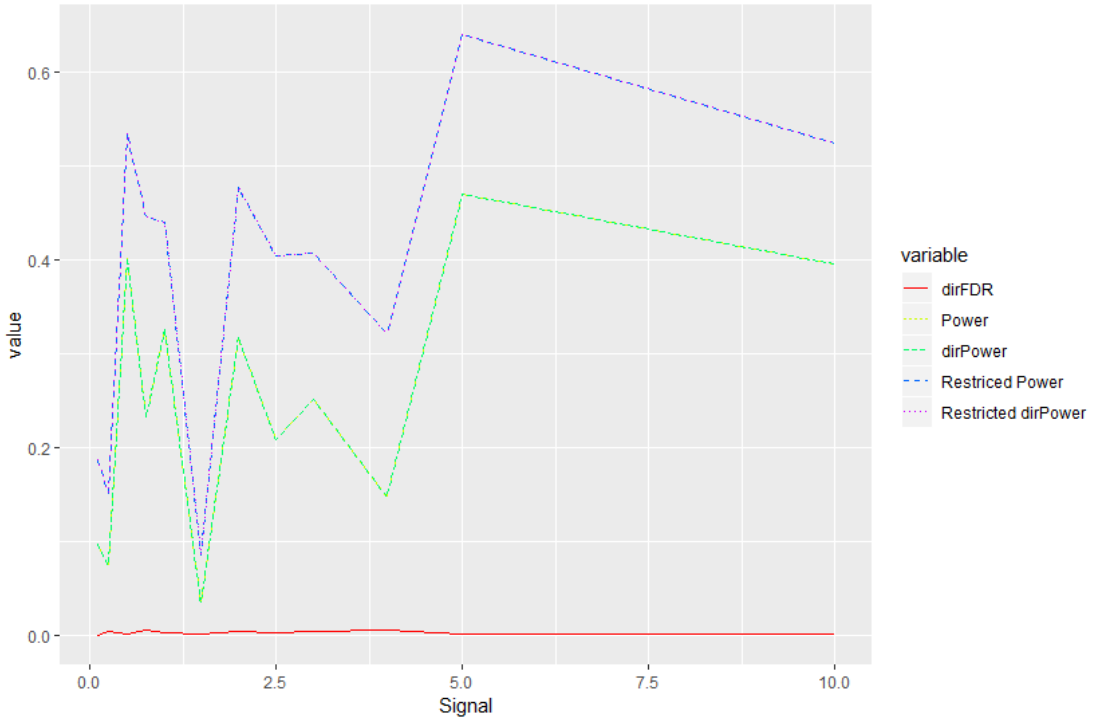

Figure 4: dirFDR and various powers as a function of true parameter magnitude.

measures are somewhat elevated. This should not be surprising, as the reduced amount of correlation between some of the variables will lead to better screening and selection by the LASSO. Note again, however, that the power is not monotonic with the magnitude of the true parameters.

To test the assertion that the normality of $\mathbf{X}$ is only used to ensure that the residuals for the partial regression performed after initial screening are still normally distributed in the proof of theorem 3 in [2], we use a non-normal distribution to generate the design matrix. Specifically, we use the multivariate log-normal distribution to generate $\mathbf{X}$. As this distribution is sub-exponential, the heavier tails should provide somewhat of a test to the validity of the assertion that the procedure is robuse to non-normality. The noise $\boldsymbol{\epsilon}$ is still normally distributed to generate the response $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, as this is an assumption of the linear regression model. We use the `rlmvnorm` function in the `dmutate` package in `R` to draw random samples from a multivariate log-normal distribution. The covariance matrix we use for the distribution is generated in the same way as above. As indicated by the plot in Figure 6, the general performance is generally poorer when compared to the previous two simulations. The scale along the y-axis shows that while the dirFDR does clearly remain below the chosen level of 0.1, all four measures of power here never exceed that level either. The compounded effects of a covariance matrix with a high average value for its off-diagonal entries and the heavier tails of a sub-exponential distribution such as the log-normal provide a formidable challenge for any

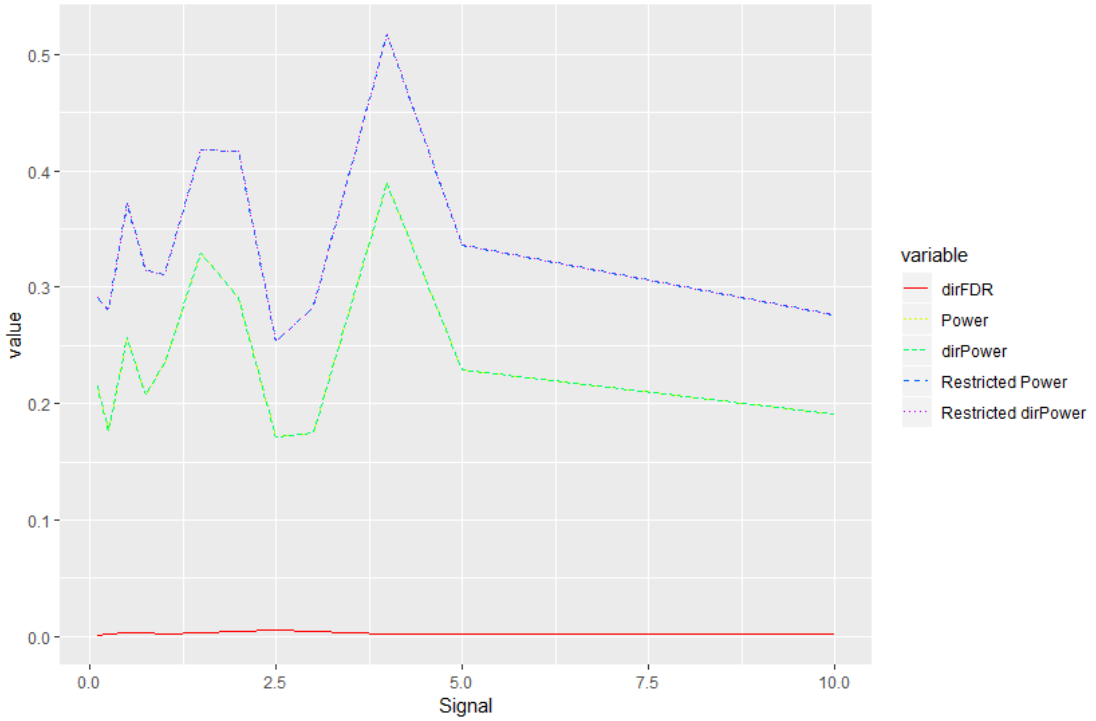

Figure 5: dirFDR and various powers as a function of true parameter magnitude, using the weighted sum covariance matrix

variable selection procedure, and here we see the knockoff method does indeed run into difficulty. In order to see if the log-normal distribution may be responsible for more of the difficulty, we re-ran the simulation using a covariance matrix that was closer to being a diagonal matrix by using a $\mathbf{W}$ that only draws entries from the symmetric interval $[-10, 10]$ and setting the number of columns k of $\mathbf{W}$ to be 1000, so that the rank of $\mathbf{WW}'$ is not deficient.

Figure 7 indicates that there is only a slight improvement in power, if any at all. Finally, we generate a design matrix from a multivariate log-normal distribution with identity covariance matrix. As Figure 8 shows, while the performance in terms of power is somewhat better here than the other two log-normal simulations, we still have substantially worse performance than when the design matrix comes from a normal distribution. This suggests that designs from a sub-exponential design, or at least a log-normal, with its heavier tails than the normal distribution, do limit the ability of the knockoff filter to make discoveries. Of course, the dirFDR is maintained below the chosen level of 0.1 in all cases, so the knockoff filter maintains the ability to guarantee that its discoveries are not false, even if these discoveries do become much fewer under a non-normal design.

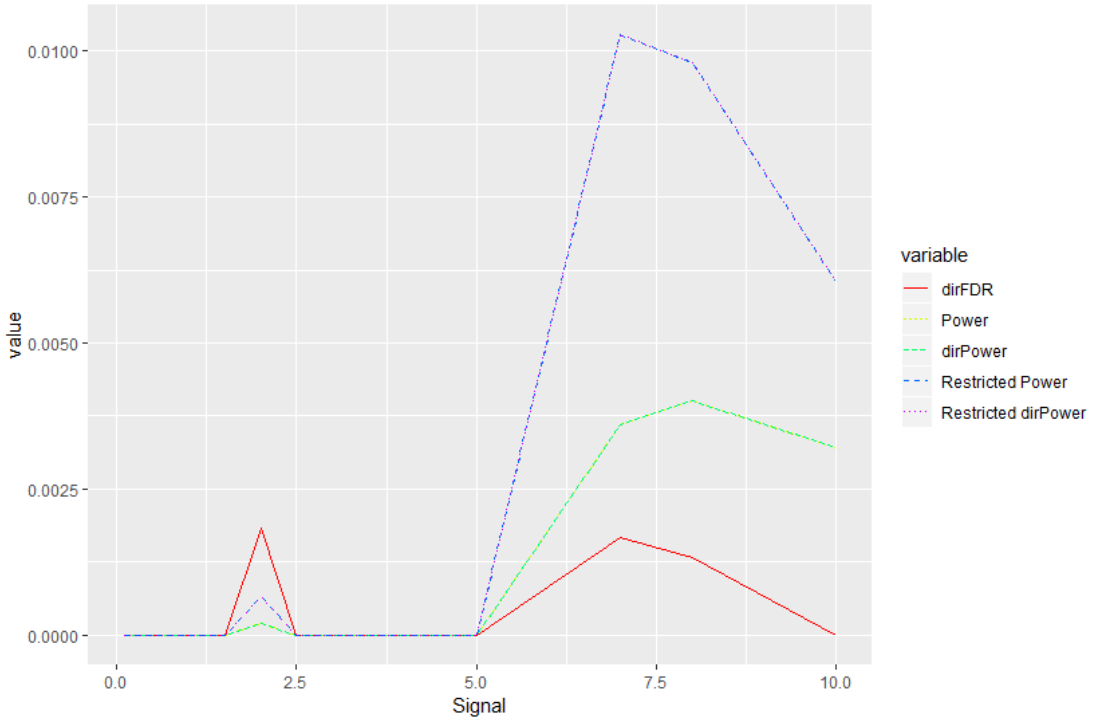

Figure 6: dirFDR and various powers as a function of true parameter magnitude, design matrix generated using log-normal distribution

4 Theoretical Guarantees

The extension of the knockoff filter of [1] to the case $p > n$ in [2] utilizes two generalizations. The first generalization is from control of FDR to control of FDR_{dir}. The second generalization requires guaranteeing this control is maintained even if an initial stage of model reduction is performed before applying the knockoff filter. As mentioned in section 2, in order to prove these results for directional control of the knockoff filter, assumptions of normality are made on the observations in [2]. These distributional assumptions on $\mathbf{y}$ or $\mathbf{X}$ are considerably stronger than the typical normality assumptions on the residuals $\boldsymbol{\epsilon}$ used to justify the use of linear regression.
There are three theorems demonstrating these generalizations in [2].

- **Theorem 1** shows that, under the assumption that $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, the standard knockoff filter developed in [1] controls FDR_{dir} when using $\widehat{\text{sign}}_j = \text{sign}\big((\mathbf{X}_j - \widetilde{\mathbf{X}}_j)^T\mathbf{y}\big)$ as the estimated direction of effect for feature j. Note that this theorem applies in the same low-dimensional realm of $n \geq 2p$, and so shows that the original knockoff filter is more powerful than shown in [1].

- **Theorem 2** again assumes that $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ and uses the same directional estimator $\widehat{\text{sign}}_j$ as in theorem 1. Here, however, we are in the high-dimensional

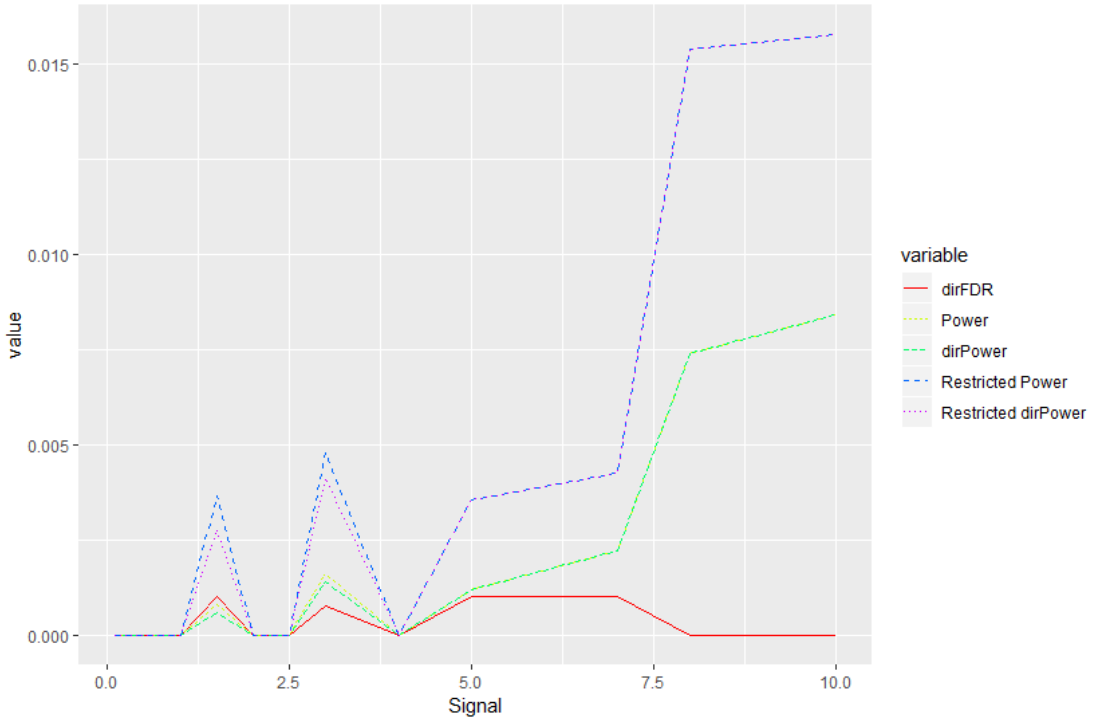

Figure 7: dirFDR and various powers as a function of true parameter magnitude, design matrix generated using log-normal distribution and 'closer-to-diagonal' covariance matrix

setting $p > n$, and thus must use an initial screening step before applying the knockoff filter. Theorem 2 asserts that the knockoff filter controls FDR_{dir} even after model reduction by some initial variable screening is performed – conditional on a *sure screening* event

$$\mathcal{E} = \{\text{support}(\boldsymbol{\beta}) \subset \widehat{S}_0 \text{ and } |\widehat{S}_0| \leq {}^{n_1}\!/_2\}. \tag{5}$$

Note that $\mathbb{1}_{\{\mathcal{E}\}}$ is a function of $\mathbf{y}^{(0)}$ for fixed $\mathbf{X}$, and so we are implicitly conditioning on the subsample used during preliminary screening. The result is

$$\mathbb{E}[\text{FDR}_{\text{dir}} \,|\, \mathcal{E}] \leq q \tag{6}$$

- **Theorem 3** extends these results to directional control of the linear regression parameters. Here the knockoff filter is employed with data recycling after initial screening. This, however, is demonstrated under the stronger assumption that the rows of $\mathbf{X}$, denoted $\mathbf{X}_{[i]}$ for $1 \leq i \leq n$, are i.i.d. samples from a multivariate Gaussian distribution,

$$\mathbf{X}_{[\mathbf{i}]} \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\nu}, \boldsymbol{\Psi}) \tag{7}$$

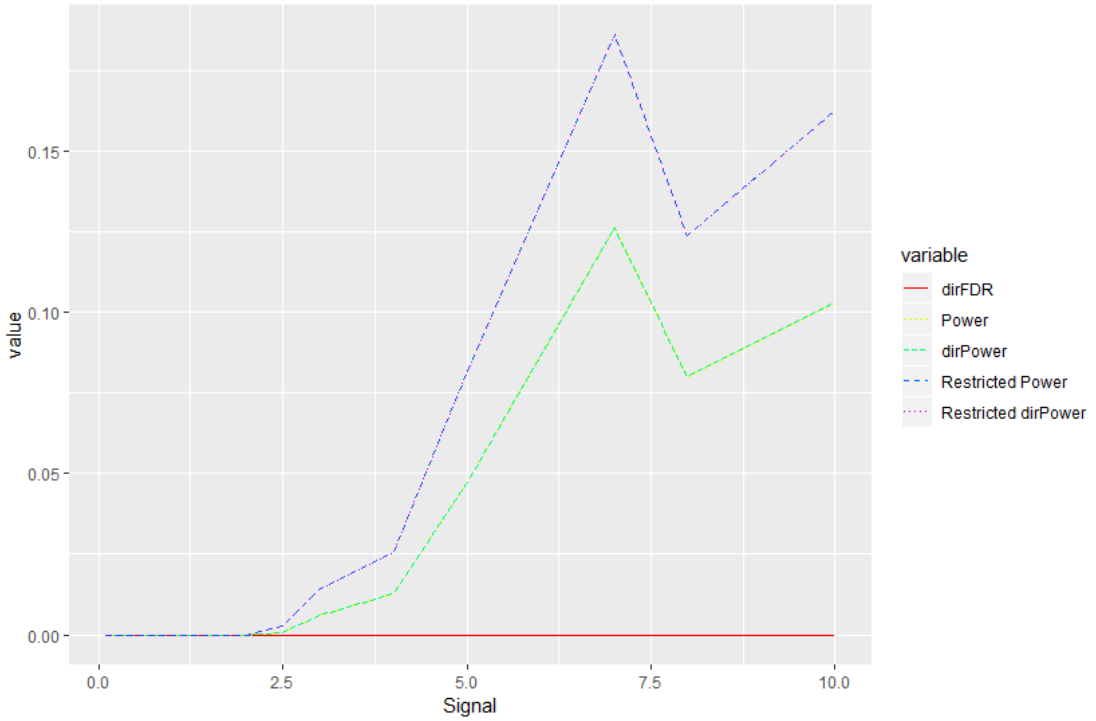

Figure 8: dirFDR and various powers as a function of true parameter magnitude, design matrix generated using log-normal distribution and 'closer-to-diagonal' covariance matrix

where $\boldsymbol{\nu} \in \mathbb{R}^p$ and $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$ are unknown. The directional control guaranteed here is also modified to refer to the *partial* regression coefficients of the reduced set $\mathbf{X}_{\widehat{S}_0}$ *expected* after initial screening on $(\mathbf{X}^{(0)}, \mathbf{y}^{(0)})$:

$$\boldsymbol{\beta}^{\text{partial}} = \left[\mathbf{X}^{(1)T}_{\widehat{S}_0}\mathbf{X}^{(1)}_{\widehat{S}_0}\right]^{-1}\mathbf{X}^{(1)T}_{\widehat{S}_0}\mathbb{E}\left[\mathbf{y}^{(1)} \,|\, \mathbf{X}^{(1)}_{\widehat{S}_0}; \mathbf{X}^{(0)}, \mathbf{y}^{(0)}\right] \tag{8}$$

As in theorem 2, conditioning is required on the subsample used for initial screening and the resulting model selection performed as a result. These conditioning arguments place the guarantees of the knockoff filter in high-dimensions more firmly in the field of selective inference [5].

Observe that in general we may have $\text{sign}(\beta_j) \neq \text{sign}(\beta_j^{\text{partial}})$. We are still tracking meaningful directional errors, however, as $\text{sign}(\beta_j^{\text{partial}})$ is the sign of the *true partial regression* coefficient for feature j if only the features in $\widehat{S}_0$ are used. While these may not be the same directional effects for the model using the true set S_0, using these partial regression coefficients mitigates the possibility that some relevant features were either dropped by initial screening or not even included in the observed set S. Thus the method still yields useful information. As mentioned by the authors, the normality assumption (7) is used so that their proof readily

reduces any errors of omission in selecting $\widehat{S}_0$ to a type of omitted variable bias familiar from standard linear regression, and a similarly recognizable change of variance for the estimate $\boldsymbol{\beta}^{\text{partial}}$ depending on the covariances of the selected $\widehat{S}_0$ and omitted $\widehat{S}_0^c$ variables.

Theorem 3 is the overall main result, at least from a practical perspective. As the knockoff filter is intended to find relevant features while placing a ceiling on the FDR_{dir}, this result most directly describes the theoretical guarantees desired to hold in practice. For this reason, the conditional nature of the statement is of major import. Both the efficacy of the selection procedure and the subsample $(\mathbf{X}^{(0)}, \mathbf{y}^{(0)})$ used for variable screening may have a potent impact on the results of the knockoff filter. Some form of cross-validation or bootstrap sampling may serve to diminish potential negative impacts incurred in this screening stage, if appropriate for the application [7]. This also illustrates the inapplicability of this form of the knockoff filter to problems of causal inference [2]. The guarantees demonstrated are for $\boldsymbol{\beta}^{\text{partial}}$ relative to $\widehat{S}_0$. Causal inference in this context would be concerned with guarantees regarding S_0 and $\boldsymbol{\beta}_{S_0}$, which as mentioned, are deferred to the selection procedure and sample used for variable screening (and the inclusion of any causal variable in the full data matrix $\mathbf{X}$ whenever a confounder for that causal variable is included).

5 Discussion

While the methods developed in [2] to extend the knockoff filter to the case where $p > n$ do represent a significant improvement, there are still some obvious limitations present in the overall knockoff pipeline as described. Perhaps the most readily apparent is that the procedure as formulated is tailored specifically to linear regression models of the form in equation (1). A recent reformulation of the knockoff method using a distributional rather than geometric construction of the false variables addresses this issue and extends the knockoff framework to generalized linear models, with no restriction on the relative magnitudes of p and n [4]. Further, these 'model-X knockoffs' are claimed to be robust to errors in estimating the distribution of $\mathbf{X}$ when generating $\widetilde{\mathbf{X}}$, in the sense that FDR control is reasonably maintained. If this leads to increased power while maintaining a controlled FDR when dealing with non-Gaussian data, this would substantially extend the applicability of the knockoff filter. As we saw in our simulations, a distribution with heavier tails such as a log-normal greatly limits the power of the knockoff filter to make discoveries, even if the FDR is maintained.

Another point that may be of concern is the error measure used by the knockoff filter. Both FDR and FDR_{dir} are defined as expected values of false discovery proportions. Depending on the application, this may not be the most salient performance measure. One alternative is the *k-familywise error rate* (k-FWER), the probability of making at least k false discoveries:

$$\mathbb{P}\big(|j : \widehat{\beta}_j \neq 0 \text{ and } \beta_j = 0| \ \geq k\big). \tag{9}$$

The knockoff methodology can be extended to control this error rate, as well as other generalized type I error rates [3].
In certain contexts, the knockoff procedure in high-dimensions controlling $\mathrm{FDR}_{\mathrm{dir}}$ is less conservative relative to the user-selected level q, and more powerful in detecting true signals when compared to earlier methods for controlling the false discovery rate. Including data recycling further reduces conservatism and increases statistical power. These properties are also somewhat more adaptive to the relative strengths of the noise in the observations and the signals from the true variables compared to other selective inference methods [2]. Overall, when applied in the appropriate contexts, the knockoff filter can produce good results. When applied in inappropriate contexts, however, it performs poorly - like any misapplied tool.

References

[1] Barber, Rina Foygel, and Emmanuel J. Candès. "Controlling the false discovery rate via knockoffs." *The Annals of Statistics* 43, no. 5 (2015): 2055-2085.

[2] Barber, Rina Foygel, and Emmanuel J. Candès. "A knockoff filter for high-dimensional selective inference." *arXiv preprint arXiv:1602.03574v3* (2018).

[3] Janson, Lucas, and Weijie Su. "Familywise error rate control via knockoffs." *Electronic Journal of Statistics* 10, no. 1 (2016): 960-975.

[4] Candes, Emmanuel, Yingying Fan, Lucas Janson, and Jinchi Lv. "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80, no. 3 (2018): 551-577.

[5] Taylor, Jonathan E. "A Selective Survey of Selective Inference." *Proc. Int. Cong. of Math. – 2018*. Rio de Janeiro, Vol. 3 (3005–3024)

[6] Zou, Hui. "The adaptive lasso and its oracle properties." *Journal of the American statistical association* 101, no. 476 (2006): 1418-1429.

[7] Wang, Sijian, Bin Nan, Saharon Rosset, and Ji Zhu. "Random lasso." *The annals of applied statistics* 5, no. 1 (2011): 468.